

Asymptotic Properties of a Branching-type Overloaded Polling Network

Zaiming Liu^a, Yuejiao Wang^a, Yuqing Chu^{b,*}, Yingqiu Li^c

^a*Department of Mathematics and Statistics, Central South University, Changsha 410083, Hunan, PR China*

^b*School of Science, Wuhan University of Technology, Wuhan 430070, Hubei, PR China*

^c*School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha 410004, Hunan, PR China*

Abstract

In this paper, we consider an N -queue overloaded polling network attended by a single cyclically roving server. Upon the completion of his service, a customer is either routed to another queue or leaves the system. All the switches are instantaneous and random multi-gated service discipline is employed within each queue. With the asymptotic theorem of multi-type branching processes and the exhaustiveness of service discipline, the fluid asymptotic process of the scaled joint queue length process is investigated. The fluid limit is of the similar shape with that of the polling system without rerouting policy, which allows us to optimize the gating indexes. Additionally, a stochastic simulation is undertaken to demonstrate the fluid limit and the optimization of the gating indexes to minimize the total population is considered.

Keywords:

Polling network, Overloaded, Multi-type branching process, Exhaustiveness, Scaled joint queue length, Stochastic simulation

*Corresponding author

Email addresses: math_lzm@csu.edu.cn (Zaiming Liu), wangyuejiaohujing@163.com (Yuejiao Wang), chuyuqing@whut.edu.cn (Yuqing Chu), liyq-2001@163.com (Yingqiu Li)

1. Introduction

A typical polling system consists of a number of queues and a single server that visits the queues in a fixed order. In recent years, polling systems have become a fascinating area of research due to their wide applications in production-inventory system, air and railway transportation, the public health system, maintenance system, computer-communication system and flexible manufacturing system (see [1, 2] for overviews). Along with that, a wide range of polling models have emerged.

In this paper, we consider a cyclic N -queue $(Q_1, \dots, Q_N, N \geq 2)$ polling system with random multi-gated service discipline within each queue and customer re-routing policies: after completing service at Q_i , a customer is either routed to Q_j with probability $p_{i,j}$ or leaves the system with probability $p_{i,0}$. The possibility for re-routing of customers further enhances the already-extensive modeling capabilities of polling models, since in many applications, customers require service at more than one facility of the system. Actually, the models of customer re-routing arise naturally in various models of computer, communication and robotic systems (see for example [3, 4, 5]). One obvious example is a local area network in which terminals are interconnected in either a physical or logical structure (see [6]).

In the vast majority of papers that have appeared on polling models, it is almost invariably assumed that the system is stable and the stable performance measures are then concerned. With the advent of the era of Internet+, the study of critically or strictly super-critically loaded polling systems is vigorously pioneered due to the overloaded Internet channel or online shopping orders.

The heavy traffic ($\rho \rightarrow 1$, ρ is the load of the system) behaviors have gained an ascending attention in the last two decades pioneered by Coffman et al. [7, 8]. By utilizing the connection with multi-type branching process, van der Mei [9] considered a unifying theory on branching-type polling models under heavy-traffic assumptions. In the similar way, Boon et al. [3] discussed the heavy-traffic asymptotic behaviors of a gated polling system with customer re-route policy. Furthermore, Liu et al. [10] extended the results in [3] to the analogous system with a general branching-type service policy in the same form. As for a non-branching type polling system, one example is that Liu et al. [11] investigated the heavy-traffic behavior of a priority polling system consisting of three M/M/1 queues with threshold policy and proved that the scaled queue-length of the critically loaded queue is

exponentially distributed, independent of that of the stable queues.

The study of overloaded ($\rho > 1$) service system is important to control or predict how fast it blows up over time. However, hardly any attention has been given to the overloaded polling system. The few literature refers to [12, 13, 14, 15]. By using measure-valued state descriptor, Puha et al. [12] proved that the overloaded $GI/GI/1$ processor sharing queues converge in distribution to supercritical fluid models and a fluid limit result is proved as first order approximations to overloaded processor sharing queues. Using both fluid and diffusion limits, Jennings et al. [14] showed that the virtual waiting time process of an overloaded Multi-class FIFO(first-in-first-out) queue with abandonments converges to a limiting deterministic fluid process. Instead Remerova et al. [15] showed the fluid asymptotic process for the joint queue length process on an overloaded branching-type cyclic polling system by using the asymptotic properties of multi-type branching processes.

In the present paper, we dedicate to the investigation of the overloaded asymptotic fluid process of the scaled joint queue length process. The fluid model associated with heavily loaded polling network is contained in [10]. The work carried out here is a natural progression from [10] and a natural extension of [15]. due to the exhaustiveness, the fluid process here has the same shape with that in [15]. For the linear shape and constant growth rate property of the fluid limit, we could optimize the system through minimizing the total queue length by choosing appropriate parameters (gating indexed and rerouting probabilities for example).

The rest of the paper is organized as follows. In Section 2, we describe precisely the polling model. In Sections 3 and 4, we introduce the supercritical property of multi-type branching process and construct the multi-branching process associated with our polling model respectively. Section 5 provides the main results, where Lemmas 4.2 and 4.4 are proved in Section 6 and Theorems 5.1 and 5.2 are proved in Section 7. Section 8 discusses some numerical issues including the stochastic simulation to test Theorems 5.1 and 5.2 and the optimization of gating indexes to minimize the average growth rate of the total population. Section 9 concludes and provides an outlook on potential further research of our paper.

2. Model description

Consider an asymmetric cyclic polling model that consists of $N \geq 2$ queues, Q_1, \dots, Q_N , and a single server that visits the queues in a cyclic order. Customers arrive at Q_i according to a Poisson process $E_i(\cdot)$ with rate λ_i . The service time of each customer at Q_i is a random variable B_i with finite mean value $\mathbb{E}B_i = 1/\mu_i$. The service discipline at Q_i is multi-gated with gating index of a random variable $\kappa_i \in \mathbb{N} \cup \{\infty\}$ (denoted by κ_i -gated, see [15]). The interarrival times, the service times and the gating indices (for different queues and for different visits) are assumed to be mutually independent.

Upon completion of service at $Q_i, i = 1, \dots, N$, a customer is either routed to $Q_j, j = 1, \dots, N$ with probability $p_{i,j}$ or leaves the system with probability $p_{i,0}$, where

$$\sum_{i=1}^N p_{i,0} > 0 \quad \text{and} \quad \sum_{j=0}^N p_{i,j} = 1.$$

We assume that all the switches of customers or servers between queues are instantaneous. Additionally, when the system becomes empty, the server travels a full cycle and subsequently stops right before Q_1 until a new arrival occurs and then cycles along the queues to serve that customer.

The total arrival rate at Q_i is denoted by γ_i , which is the unique solution of the following set of linear equation ([4]):

$$\gamma_i = \lambda_i + \sum_{j=1}^N \gamma_j p_{j,i} \quad i = 1, \dots, N.$$

The offered load to Q_i equals to $\rho_i = \gamma_i/\mu_i$ and the total load equals to $\rho = \sum_{i=1}^N \rho_i$. Furthermore, we need two more assumptions.

Assumption 1. For all $i = 1, \dots, N$, $\rho_i < 1$, and $\rho > 1$.

Assumption 2. For all $i = 1, \dots, N$, $\mathbb{E}B_i \log B_i < \infty$.

Throughout the paper, we focus our attention on the overloaded behavior of the queue length process $\mathbf{X}(\cdot) = (X_1, \dots, X_N)(\cdot)$, where $X_i(t)$ is the number of customers at Q_i at time t .

For simplicity, we adopt the following notations.

1. Let $\mathbb{N} = \{0, 1, \dots\}$ and $\mathbb{N}^+ = \{1, 2, \dots\}$.
2. The vector with all coordinates equal to 0 is denoted by $\mathbf{0}$, with all coordinates equal to 1 by $\mathbf{1}$, and with coordinate i equals to 1 and the other coordinates equal to 0 by \mathbf{e}_i .
3. For vectors $\mathbf{x} = (x_1, \dots, x_N)$ and $\mathbf{y} = (y_1, \dots, y_N)$, define the following operations:
 - coordinate-wise product: $\mathbf{x} \times \mathbf{y} = (x_1 y_1, \dots, x_N y_N)$;
 - power: if all $x_i > 0$, $1 \leq i \leq N$, then $\mathbf{x}^{\mathbf{y}} = \prod_{i=1}^N x_i^{y_i}$;
 - binomial coefficient: if $\mathbf{y} \leq \mathbf{x}$, then $\binom{\mathbf{x}}{\mathbf{y}} = \prod_{i=1}^N \binom{x_i}{y_i} = \prod_{i=1}^N x_i! / y_i! (x_i - y_i)!$.
4. If random objects X and Y are equal in distribution, we write $X \stackrel{d}{=} Y$ and say that X is a copy of Y .

3. Multi-type branching process

We consider a general multi-type branching process (MTBP) with N particle types, denoted by $\mathbf{Z}_n = (Z_n^{(1)}, \dots, Z_n^{(N)})$, $n \in \mathbb{N}$, where $Z_n^{(i)}$ is the number of type- i particles in the n th generation for $i = 1, \dots, N$, $n \in \mathbb{N}$.

- Define the immigration distribution by

$$G(\mathbf{k}) := \mathbb{P}(\mathbf{Z}_{n+1} = \mathbf{k} | \mathbf{Z}_n = \mathbf{0}), \quad \mathbf{k} \in \mathbb{N}^N.$$

- Let $\mathbf{M} = \{m_{i,j}\}_{i,j=1}^N$ be the mean offspring matrix, where $m_{i,j}$ is the expected number of type j offspring of a single type i particle in one generation.
- Let the vectors $\mathbf{u} = (u_1, \dots, u_N)$ and $\mathbf{v} = (v_1, \dots, v_N)$ be the right and left eigenvectors corresponding to the maximal real-valued, positive eigenvalue θ of \mathbf{M} , commonly referred to as the maximum eigenvalue ([16]), normalized such that $\mathbf{v}\mathbf{u}^\top = 1$.
- Let $\mathbf{q} = (q_1, \dots, q_N)$, where q_i is the probability of eventual extinction of the process initiated with a single particle of type i , i.e.,

$$q_i = \mathbb{P}(\mathbf{Z}_n = \mathbf{0} \text{ for some } n | \mathbf{Z}_0 = \mathbf{e}_i).$$

It follows that the auxiliary process $\{\mathbf{Z}^{(n)}\}_{n \in \mathbb{N}}$ has the following asymptotics.

Proposition 3.1. ([16]) *Given $\mathbf{Z}^{(0)} = \mathbf{e}_i$,*

$$\frac{\mathbf{Z}^{(n)}}{\theta^n} \rightarrow \xi_i \mathbf{v} \quad \text{almost surely (a.s.)} \quad \text{as } n \rightarrow \infty,$$

where the distribution of the random variable ξ_i has a point mass $q_i < 1$ at 0 and a continuous density function on $(0, \infty)$ with $\mathbb{E}\xi_i = u_i$.

4. Branching property of multi-gated polling system

To start with, define $t^{(n)}$ as the time point that the server reaches right before Q_1 for the n th time and $t_i^{(n)}$ as the time point that the server reaches Q_i for the n th time ($n \in \mathbb{N}^+, i = 1, 2, \dots, N$), then

1. $t^{(n)} \leq t_1^{(n)} \leq \dots \leq t_{N+1}^{(n)} = t^{(n+1)}$;
2. If the system is empty at $t^{(n)}$, then the interval $[t^{(n)}, t_1^{(n)})$ is the period of waiting until the first arrival, otherwise $t^{(n)} = t_1^{(n)}$;
3. The interval $[t_i^{(n)}, t_{i+1}^{(n)})$ is the visit time at Q_i following $t^{(n)}$, with $t_i^{(n)} = t_{i+1}^{(n)}$ if Q_i is empty.

Typically, we assume $t^{(1)} = 0$, which means the system is empty at $t = 0$.

Branching Property[17] If the server arrives at Q_i to find k_i customers there, then during the course of the server's visit, each of these k_i customers will effectively be replaced in an i.i.d. manner by a random population having probability generating function (p.g.f) $h_i(z_1, z_2, \dots, z_N)$, which can be any N -dimensional p.g.f..

By Resing [17], branching property implies that the queue length sequence $\{\mathbf{X}(t^{(n)})\}_{n \in \mathbb{N}}$ forms a multi-type branching process with immigration in state $\mathbf{0}$. Then the probability for the process $\{\mathbf{X}(t^{(n)})\}_{n \in \mathbb{N}}$ to return to $\mathbf{0}$ is given by

$$q_G := \sum_{\mathbf{k} \in \mathbb{N}^N} G(\mathbf{k}) \mathbf{q}^{\mathbf{k}}.$$

Subsequently, we give some notations associated with the branching-type polling system.

- Define $\check{\mathbf{L}}_i = (\check{L}_{i,1}, \dots, \check{L}_{i,N})$ as the visit offspring of a customer at Q_i , which equals in distribution to $\mathbf{X}(t_{i+1}^{(n)})$ given that $\mathbf{X}(t_i^{(n)}) = \mathbf{e}_i$ (its distribution does not depend on n).
- Define $\mathbf{L}_i := (L_{i,1}, \dots, L_{i,N})$ as the session offspring of a customer at Q_i , which equals in distribution to $\mathbf{X}(t^{(n+1)})$ given that $\mathbf{X}(t^{(n)}) = \mathbf{e}_i$ (its distribution does not depend on n).
- Denote the mean visit offspring of a customer at Q_i and the mean session offspring by $\check{\mathbf{m}}_i = (\check{m}_{i,1}, \dots, \check{m}_{i,N}) = \mathbb{E}\check{\mathbf{L}}_i$ and $\mathbf{m}_i = (m_{i,1}, \dots, m_{i,N}) = \mathbb{E}\mathbf{L}_i$, respectively.
- Define T_i as the visit duration at Q_i which equals in distribution to $t_{i+1}^{(n)} - t_i^{(n)}$ given that $X_i(t_i^{(n)}) = 1$.

Then the immigration distribution is given by

$$G(\mathbf{k}) = \frac{\sum_{i=1}^N \lambda_i \mathbb{P}(\mathbf{L}_i = \mathbf{k})}{\sum_{i=1}^N \lambda_i}, \quad \mathbf{k} \in \mathbb{N}^N. \quad (1)$$

Before proceeding further, we give a lemma on the exhaustiveness f_i of the service discipline at Q_i , which is defined by (see [18])

$$f_i = 1 - \frac{\partial}{\partial z_i} h_i(z_1, z_2, \dots, z_N) \big|_{\mathbf{z}=\mathbf{1}} = 1 - \mathbb{E}\check{L}_{i,i}.$$

Virtually, it has an appealing interpretation: during the course of the server's visit at Q_i , each customer present at the start of the visit to Q_i will be replaced by a number of customers with mean $1 - f_i$ at the end of the visit to Q_i .

Lemma 4.1. *In our model, for Q_i , we have*

$$f_i = 1 - \mathbb{E}\left(\frac{\lambda_i}{\mu_i} + p_{i,i}\right)^{\kappa_i}, \quad (2)$$

$$t_i := \mathbb{E}T_i = \frac{1 - \mathbb{E}\left(\frac{\lambda_i}{\mu_i} + p_{i,i}\right)^{\kappa_i}}{\mu_i(1 - \frac{\lambda_i}{\mu_i} - p_{i,i})}. \quad (3)$$

Let B_i^E and Y_i be the total service time and the number of services of a customer in Q_i before he is either routed to $Q_j, j \neq i$ or leaves the system. Obviously, $\mathbb{P}(Y_i = n) = p_{i,i}^{n-1}(1 - p_{i,i}), (n = 1, 2, \dots)$ and

$$b_i^E = \mathbb{E}B_i^E = \mathbb{E} \sum_{j=1}^{Y_i} B_{i,j} = \mathbb{E}Y_i \mathbb{E}B_i = \frac{1}{\mu_i(1 - p_{i,i})},$$

where $\{B_{i,j}, j = 1, 2, \dots\}$ are i.i.d. copies of B_i . By Lemma 1 in [10], the mean offspring matrix \mathbf{M} is given in the following lemma.

Lemma 4.2. *For the cyclic branching-type polling system, the mean matrix \mathbf{M} is given by*

$$\mathbf{M} = \mathbf{M}_1 \dots \mathbf{M}_N,$$

where $\mathbf{M}_k = \left(m_{i,j}^{(k)}\right)$ and

$$m_{i,j}^{(k)} = \begin{cases} \delta_{\{i=j\}}, & i \neq k, \\ 1 - f_i, & i = k = j, \\ f_i \varphi_i(\mu_i p_{i,j} + \lambda_j), & i = k \neq j, \end{cases}$$

where δ_F denotes the indicator function on F and $\varphi_i = \frac{b_i^E}{1 - \lambda_i b_i^E}$.

Actually, \mathbf{M}_k is the mean session offspring during the visit time on Q_k . Hence, for all i ,

$$\tilde{m}_{i,j} = m_{i,j}^{(i)} = \begin{cases} 1 - f_i, & i = j, \\ f_i \varphi_i(\mu_i p_{i,j} + \lambda_j), & i \neq j. \end{cases}$$

Thus, for the $m_{i,j}$, we also have the following recursive formula:

$$m_{N,j} = \tilde{m}_{N,j}, \quad \text{for all } j,$$

and, for $i = 1, \dots, N-1$, \mathbf{m}_i is computed via \mathbf{m}_{i+1} ,

$$m_{i,j} = \tilde{m}_{i,j} \delta_{\{i \geq j\}} + \sum_{k=i+1}^N \tilde{m}_{i,k}, \quad \text{for all } j.$$

By the properties of the maximal eigenvalue θ of the mean matrix \mathbf{M} , the following Lemma 4.3 construct the connection between extinction probability and the maximal eigenvalue θ in branching-type polling systems. For notational convenience, denote $\theta(\rho)$ by θ .

Lemma 4.3. *For the Perron-Frobenius eigenvalue θ and the extinction probabilities q_i , we have $\theta > 1$ and $q_i < 1$ for all i . By the latter, $q_G < 1$, too.*

It follows that the auxiliary process $\{\mathbf{X}(t^{(n)})\}_{n \in \mathbb{N}}$ has the following asymptotics.

Proposition 4.1. *If the first arriving customer arrives at Q_i after $t = 0$, then*

$$\frac{\mathbf{X}(t^{(n)})}{\theta^n} \rightarrow \xi_i \mathbf{v} \quad \text{almost surely (a.s.)} \quad \text{as } n \rightarrow \infty,$$

where the distribution of the random variable ξ_i has a jump of magnitude $q_i < 1$ at 0 and a continuous density function on $(0, \infty)$ and $\mathbb{E}\xi_i = u_i$.

Lemma 4.4 below considers the finiteness of the corresponding moments for the offspring distribution of the multi-type branching processes $\{\mathbf{Q}(t^{(n)})\}_{n \in \mathbb{N}}$, which guarantees the non-degenerate of the random variable ξ_i .

Lemma 4.4. *For all i and j , $\mathbb{E}L_{i,j} \log L_{i,j} < \infty$, where $0 \log 0 := 0$ by convention.*

5. Fluid limit

To give the main results, two more notations are needed.

- Let \bar{B}_i be the total service time of a customer arriving at Q_i from outside, $\bar{c}_i = \mathbb{E}\bar{B}_i$ and A_i be the position of a customer after completion of service at Q_i , for $i = 1, \dots, N$, i.e.,

$$A_i = \begin{cases} j, & \text{after receiving service at } Q_i, \text{ a customer is routed to } Q_j; \\ 0, & \text{after receiving service at } Q_i, \text{ a customer leaves the system.} \end{cases}$$

Then $\mathbb{P}(A_i = j) = p_{i,j}$, $j = 0, 1, \dots, N$. By the law of total probability, we have

$$\begin{aligned} \bar{c}_i &= \mathbb{E}\bar{B}_i = \mathbb{E}(\bar{B}_i | A_i = 0) \mathbb{P}(A_i = 0) + \sum_{j=1}^N \mathbb{E}(\bar{B}_i | A_i = j) \mathbb{P}(A_i = j) \\ &= \mathbb{E}B_i + \sum_{j=1}^N p_{i,j} \mathbb{E}\bar{B}_j = 1/\mu_i + \sum_{j=1}^N p_{i,j} \bar{c}_j. \end{aligned}$$

It is also easy to deduce that $\rho = \sum_{i=1}^N \lambda_i \bar{c}_i$.

- For $n \in \mathbb{N}$, let

$$\eta_n := \begin{cases} \min\{k : t^{(k)} \geq \theta^n\}, & \text{if } n \geq 0; \\ 0, & \text{if } n < 0. \end{cases}$$

Theorem 5.1. *There exist constants $\bar{b}_i \in (0, \infty)$ and $\bar{\mathbf{a}}_i = (\bar{a}_{i,1}, \dots, \bar{a}_{i,N}) \in [0, \infty)^N, i = 1, \dots, N+1$, and a random variable ξ with values in $[1, \theta)$ such that, for all $k \in \mathbb{N}$ and i ,*

$$\frac{t_i^{(\eta_n+k)}}{\theta^n} \rightarrow \theta^k \bar{b}_i \xi \quad \text{and} \quad \frac{X(t_i^{(\eta_n+k)})}{\theta^n} \rightarrow \xi \theta^k \bar{\mathbf{a}}_i \quad \text{a.s. as } n \rightarrow \infty.$$

The \bar{b}_i and $\bar{\mathbf{a}}_i$ are given by

$$\bar{b}_1 = 1, \quad \bar{b}_{i+1} = \bar{b}_i + \left[\frac{v_i}{\alpha} + \lambda_i (\bar{b}_i - \bar{b}_1) + \sum_{j=1}^{i-1} p_{j,i} \mu_j (\bar{b}_{j+1} - \bar{b}_j) \right] t_i, \quad i = 1, \dots, N;$$

and for $i = 1, \dots, N$,

$$\bar{\mathbf{a}}_1 = \frac{\mathbf{v}}{\alpha}, \quad \bar{a}_{i+1,j} = \begin{cases} \bar{a}_{i,j} + [\lambda_j + \mu_i p_{i,j}] (\bar{b}_{i+1} - \bar{b}_i) & j \neq i, \\ \bar{a}_{i,i} + [\lambda_i - \mu_i (1 - p_{i,i})] (\bar{b}_{i+1} - \bar{b}_i), & j = i. \end{cases}$$

where $\alpha = \frac{\sum_{i=1}^N v_i \bar{c}_i}{\rho - 1}$. For $x \in [1, \theta)$, the distribution of ξ is given by

$$\begin{aligned} \mathbb{P}(\xi \geq x) &= \frac{1}{1 - q_G} \sum_{\substack{\mathbf{k} \in \mathbb{N}^N \\ |\mathbf{k}| \geq 1}} G(\mathbf{k}) \sum_{\substack{1 \leq \mathbf{k} \\ |\mathbf{l}| \geq 1}} \binom{\mathbf{k}}{\mathbf{l}} (1 - \mathbf{q})^{\mathbf{l}} \mathbf{q}^{\mathbf{k}-\mathbf{l}} \\ &\times \mathbb{P} \left\{ \left\{ \log_{\theta} \left(\alpha \sum_{i=1}^N \sum_{j=1}^{l_i} \xi_i^{(j)} \right) \right\} \geq \log_{\theta} x \right\} \end{aligned}$$

where $\xi_i^{(j)}, j \in \mathbb{N}^+$ are i.i.d. copies of $\xi_i \delta_{\xi_i > 0}$.

For each $n \in \mathbb{N}$, define the scaled queue length process

$$\bar{\mathbf{X}}^{(n)}(t) := \frac{\mathbf{X}(\theta^n t)}{\theta^n}, \quad t \in [0, \infty). \quad (4)$$

Theorem 5.2. *There exists a deterministic function $\bar{\mathbf{X}}(\cdot) = (\bar{X}_1, \dots, \bar{X}_N)(\cdot) \in [0, \infty)^N$ such that,*

$$\bar{\mathbf{X}}^{(n)}(\cdot) \rightarrow \xi \bar{\mathbf{X}}\left(\frac{\cdot}{\xi}\right) \quad \text{a.s.} \quad \text{as } n \rightarrow \infty,$$

uniformly on compact sets, where the random variable ξ is defined in Theorem 5.1. For all $i = 1, \dots, N$, the function $\bar{\mathbf{X}}(\cdot)$ is continuous and piecewise linear and specified by

$$\bar{\mathbf{X}}(t) = \begin{cases} 0, & \text{if } t = 0; \\ \theta^k \bar{\mathbf{a}}_i + (t - \theta^k \bar{b}_i) \boldsymbol{\lambda} + (t - \theta^k \bar{b}_i) \mu_i \mathbf{p}_i, & \text{if } t \in [\theta^k \bar{b}_i, \theta^{k+1} \bar{b}_{i+1}), \end{cases} \quad (5)$$

where $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_N]$ and $\mathbf{p}_i = [p_{i,1}, \dots, p_{i,i-1}, p_{i,i} - 1, p_{i,i+1}, \dots, p_{i,N}]$.

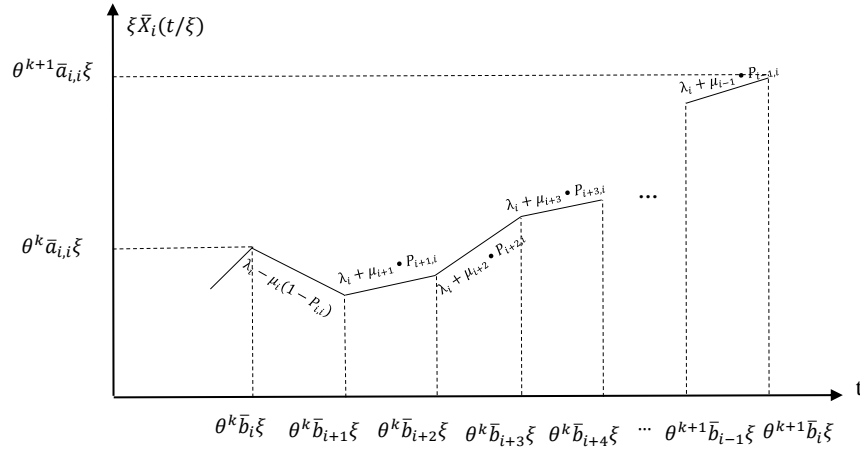


Fig. 1 Fluid limit of Q_i .

Corollary 5.1. *The limit total population $(\bar{X}_1 + \bar{X}_2 + \dots + \bar{X}_N)(\cdot)$ grows at rate*

$$(\lambda_1 + \dots + \lambda_N) - p_{i,0} \mu_i$$

when $t \in [\theta^k \bar{b}_i, \theta^{k+1} \bar{b}_{i+1})$ for all $k \in \mathbb{N}$.

According to Theorem 5.2 and Corollary 5.1, the fluid limit processes both demonstrate an oscillation waveform with increasing amplitude and cycle time forward through time and oscillate at an infinite rate when approaching zero. To be more specific, the amplitude and cycle time both

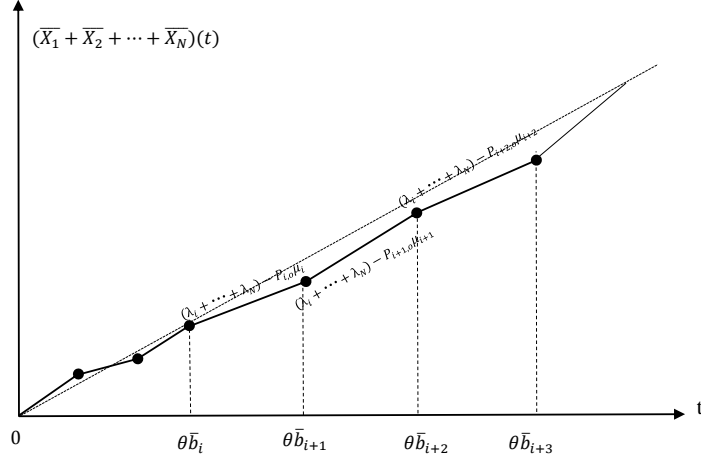


Fig. 2 Fluid limit of total population $\bar{X}_1 + \dots + \bar{X}_N$. \cdots is the average growth rate of total population.

increase by $\theta - 1$ times each cycle. Hence, the average growth rates of the fluid limit in each cycle equal to the average growth rate in the whole time. As shown in Fig.2, the average growth rate, denoted by β , can be given by

$$\sum_{i=1}^N \frac{\sum_{j=1}^N \bar{a}_{i,j} + \sum_{j=1}^N \bar{a}_{i+1,j}}{2} (\bar{b}_{i+1} - \bar{b}_i) = \int_{\bar{b}_1}^{\theta \bar{b}_1} \beta t dt,$$

which yields

$$\beta = \frac{1}{\theta^2 - 1} \left[\sum_{i=1}^N \left(\sum_{j=1}^N \bar{a}_{i,j} + \sum_{j=1}^N \bar{a}_{i+1,j} \right) (\bar{b}_{i+1} - \bar{b}_i) \right]. \quad (6)$$

By the definition of the scaled queue length process, the fluid limit could approximate the original queue length process in steady state. Furthermore, the average growth rate in (6) allows us to study the optimization problem of how to choose the gating indexes of each queue to minimize the total queue length. Because each of the queues adheres to a branching-type service discipline, we also study how to choose the exhaustiveness f_i with the same objective in mind. Here we only provide an example by utilizing the genetic algorithm to solve the optimization problem in Section 8.

6. Proof of Lemma 4.1 and Lemma 4.4

Proof of Lemma 4.1. By (4.1) and (4.2) in [10], we have $f_i = \frac{\mathbb{E}T_i}{\varphi_i}$, where φ_i is defined in Lemma 4.2. Hence, it only remains to prove (3).

For $k \in \mathbb{N} \cup \{\infty\}$, let $T_i(k)$ be the visit duration at Q_i given that the service discipline at Q_i is k -gated. Obviously, we have $\mathbb{E}T_i(\infty) = \frac{1}{\mu_i(1 - \frac{\lambda_i}{\mu_i} - p_{i,i})}$.

Since

$$T_i(0) = 0 \quad \text{and} \quad T_i(k+1) \stackrel{d}{=} B_i + \sum_{l=1}^{E_i(B_i)} T_i^{(l)}(k) + \delta_{\{A_i=i\}} T_i(k), \quad k \in \mathbb{N},$$

where $T_i^{(l)}(k), l \in \mathbb{N}^+$ are i.i.d. copies of $T_i(k)$ and the random elements $B_i, E_i(\cdot)$ and $\{T_i^{(l)}(k)\}_{l \in \mathbb{N}^+}$ are mutually independent, we get

$$\begin{aligned} \mathbb{E}T_i(k+1) &= \frac{1}{\mu_i} + \left(\frac{\lambda_i}{\mu_i} + p_{i,i}\right) \mathbb{E}T_i(k) \\ &= \frac{1}{\mu_i} \left(1 + \frac{\lambda_i}{\mu_i} + p_{i,i}\right) + \left(\frac{\lambda_i}{\mu_i} + p_{i,i}\right)^2 \mathbb{E}T_i(k-1) \\ &= \dots = \frac{1}{\mu_i} \frac{1 - \left(\frac{\lambda_i}{\mu_i} + p_{i,i}\right)^{k+1}}{1 - \frac{\lambda_i}{\mu_i} - p_{i,i}}. \end{aligned} \tag{7}$$

Therefore,

$$\mathbb{E}T_i = \sum_{k \in \mathbb{N} \cup \{\infty\}} \mathbb{P}(\kappa_i = k) \mathbb{E}T_i(k) = \frac{1 - \mathbb{E}\left(\frac{\lambda_i}{\mu_i} + p_{i,i}\right)^{\kappa_i}}{\mu_i \left(1 - \frac{\lambda_i}{\mu_i} - p_{i,i}\right)}.$$

In order to prove Lemma 4.4, we need the following lemmas.

Lemma 6.1. ([15], Lemma 4) Suppose that a function $f(\cdot) : [0, \infty) \rightarrow [0, \infty)$ is bounded in a finite interval $[0, T]$ and nondecreasing in $[T, \infty)$, and that $f(x) \rightarrow \infty$ as $x \rightarrow \infty$. Suppose also that, for some (and, hence, for all) $c > 1$,

$$\limsup_{x \rightarrow \infty} \frac{f(cx)}{f(x)} < \infty.$$

Consider an i.i.d. sequence $\{Y^{(n)}\}_{n \in \mathbb{N}^+}$ of nonnegative, non degenerate at 0 random variables, and the renewal process

$$Y(t) = \max\{n \in \mathbb{N} : \sum_{k=1}^n Y^{(k)} \leq t\}, \quad t \in [0, \infty).$$

Let τ be a nonnegative random variable which may depend on the sequence $\{Y_n\}_{n \in \mathbb{N}^+}$. Assume that $\mathbb{E}f(\tau) < \infty$. Then $\mathbb{E}f(Y(\tau))$ is finite too.

Lemma 6.2. ([15], Lemma 5) Consider a sequence $Y_{n \in \mathbb{N}^+}^{(n)}$ of nonnegative random variables that are identically distributed (but not necessarily independent), and also a \mathbb{N} -valued random variable η that does not depend on $Y_{n \in \mathbb{N}^+}^{(n)}$. If $f(\cdot) : [0, \infty) \rightarrow \mathbb{R}$ is a convex function, then

$$\mathbb{E}f\left(\sum_{k=1}^{\eta} Y^{(k)}\right) \leq \mathbb{E}f(\eta Y^{(1)}).$$

We continue proving Lemma 4.4. It suffices to show that

$$\mathbb{E}f(L_{ij}) < \infty, \quad \text{for all } i \text{ and } j,$$

where

$$f(x) = \begin{cases} 0, & x \in [0, 1]; \\ x \log x, & x \in [1, \infty). \end{cases}$$

Note that the function $f(\cdot)$ is convex in $(1, \infty)$, its derivative $\log(\cdot) + 1$ is nondecreasing, and in the other points, it is easy to verify the convexity. Note also that

$$f(xy) \leq xf(y) + yf(x), \quad x, y \in [0, \infty). \quad (8)$$

Proof of Lemma 4.4. (1) **Finiteness of $\mathbb{E}f(T_i)$.** Recall the definition of $T_i(k)$ in the proof of Lemma 4.1, we have

$$T_i(k) \uparrow T_i(\infty) \quad a.s. \quad \text{as } k \rightarrow \infty.$$

Then by the monotonicity, convexity of $f(\cdot)$ and the auxiliary Lemma 6.2 combined with (8), we have

$$\begin{aligned} \mathbb{E}f(T_i(k)) &\leq \mathbb{E}f(T_i(k+1)) = \mathbb{E}f\left(B_i + \sum_{l=1}^{E_i(B_i)} T_i^{(l)}(k) + \delta_{\{A_i=i\}} T_i(k)\right) \\ &\leq \frac{1}{3} \mathbb{E}f(3B_i) + \frac{1}{3} \mathbb{E}f\left(3 \sum_{l=1}^{E_i(B_i)} T_i^{(l)}(k)\right) + \frac{1}{3} \mathbb{E}f(3\delta_{\{A_i=i\}} T_i(k)) \end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{3}\mathbb{E}f(3B_i) + \frac{1}{3}\mathbb{E}f(3E_i(B_i)T_i^{(l)}(k)) + \frac{1}{3}\mathbb{E}f(3\delta_{\{A_i=i\}}T_i(k)) \\
&\leq \frac{1}{3}\mathbb{E}f(3B_i) + \mathbb{E}[E_i(B_i)]\mathbb{E}f(T_i^{(l)}(k)) + \frac{1}{3}\mathbb{E}T_i^{(l)}(k)\mathbb{E}f(3E_i(B_i)) \\
&\quad + \mathbb{E}\delta_{\{A_i=i\}}\mathbb{E}f(T_i(k)) + \frac{1}{3}\mathbb{E}T_i(k)\mathbb{E}f(3\delta_{\{A_i=i\}}) \\
&\leq \frac{1}{3}\mathbb{E}f(3B_i) + \left(\frac{\lambda_i}{\mu_i} + p_{i,i}\right)\mathbb{E}f(T_i(k)) + \frac{1}{3}\mathbb{E}T_i(k)[f(3) + \mathbb{E}f(3E_i(B_i))].
\end{aligned}$$

By Lemma 6.1 and (7), we have $\mathbb{E}f(3E_i(B_i)) < \infty$ and $\mathbb{E}T_i(k) \leq \frac{1}{\mu_i(1 - \frac{\lambda_i}{\mu_i} - p_{i,i})}$.

Therefore, for all $k \geq 2$, we get

$$\mathbb{E}f(T_i(k)) \leq \frac{C}{1 - \frac{\lambda_i}{\mu_i} - p_{i,i}},$$

where

$$C = \frac{1}{3}\mathbb{E}f(3B_i) + \frac{1}{3} \frac{f(3) + \mathbb{E}f(3E_i(B_i))}{\mu_i(1 - \frac{\lambda_i}{\mu_i} - p_{i,i})} < \infty.$$

(2) **Finiteness of $\mathbb{E}f(\check{L}_{i,i})$.** $\check{L}_{i,i}$ is bounded stochastically from above by the number of customers arriving from outside when the system is empty and during $T_i(\infty)$ (the visit duration at Q_i with exhaustive service policy). Therefore, $\check{L}_{i,i} < 1 + E_i(T_i(\infty))$. Hence, $\mathbb{E}f(\check{L}_{i,i}) < \mathbb{E}f(1 + E_i(T_i(\infty))) < \infty$ by Lemma 6.1.

(3) **Finiteness of $\mathbb{E}f(L_{i,j})$.** The result can be proved in the same way as in the proof of Lemma 3 in [15].

7. Proof of Theorems 5.1 and 5.2

To prove Theorems 5.1 and 5.2, we need some further notations:

1. Define the renewal processes

$$Y_i(t) = \max \left\{ n \in \mathbb{N}^+, \text{ such that } \sum_{i=1}^n B_i^{(k)} \leq t \right\},$$

where $B_i^{(k)}$ are i.i.d. copies of B_i .

2. Let $I_i(t)$ be the whole time that the server has spent at Q_i before time t , i.e.,

$$I_i(t) = \int_0^t I(\text{queue } i \text{ is in service in time } s) ds \quad t \in (0, \infty).$$

3. Define index ν by

$$\nu = \max \{n \in \mathbb{N}^+, \text{ such that } \mathbf{X}(t^{(n)}) = \mathbf{0} \text{ and } \mathbf{X}(t^{(m)}) \neq \mathbf{0} \text{ for all } m > n\}.$$

Lemma 7.1. ([15], Proposition 2) *Let a random variable Y have a finite mean value and, for each $n \in \mathbb{N}^+$, let $Y_n^{(k)}, k \in \mathbb{N}^+$ be i.i.d. copies of Y . Let $\tau_n, n \in \mathbb{N}^+$ be \mathbb{N} -valued random variables such that τ_n is independent of the sequence $\{Y_n^{(k)}\}_{k \in \mathbb{N}^+}$ for each n and $\tau_n \rightarrow \infty$ in probability as $n \rightarrow \infty$. Finally, let a sequence $\{T_n\}_{n \in \mathbb{N}^+}$ of positive numbers increase to ∞ . If there exists an a.s. finite random variable τ such that $\tau_n/T_n \rightarrow \tau$ in probability as $n \rightarrow \infty$, then*

$$\sum_{k=1}^{\tau_n} \frac{Y_n^{(k)}}{T_n} \rightarrow \tau \mathbb{E}Y \quad \text{in probability} \quad \text{as } n \rightarrow \infty.$$

Lemma 7.2. *For $i = 1, \dots, N$, there exist constants $b_i \in (0, \infty)$ and $\mathbf{a}_i = (a_{i,1}, \dots, a_{i,N}) \in [0, \infty)^N$ such that*

$$\frac{t_i^{(n)}}{\theta^n} \rightarrow b_i \xi \quad \text{and} \quad \frac{\mathbf{X}(t_i^{(n)})}{\theta^n} \rightarrow \xi \mathbf{a}_i \quad \text{a.s. as } n \rightarrow \infty.$$

The b_i 's and \mathbf{a}_i 's are specified by

$$b_1 = \frac{\sum_{i=1}^N v_i \bar{c}_i}{\rho - 1},$$

$$b_{i+1} = b_i + \left(v_i + \lambda_i(b_i - b_1) + \sum_{j=1}^{i-1} p_{j,i} \mu_j (b_{j+1} - b_j) \right) t_i, \quad i = 1, \dots, N, \quad (9)$$

and

$$\mathbf{a}_1 = \mathbf{v}, a_{i+1,j} = \begin{cases} a_{i,j} + (\lambda_j + p_{i,j} \mu_i)(b_{i+1} - b_i), & j \neq i; \\ a_{i,i} + (\lambda_i - \mu_i(1 - p_{i,i}))(b_{i+1} - b_i), & j = i. \end{cases} \quad (10)$$

Obviously, the \mathbf{a}_i 's also satisfy

$$\mathbf{a}_1 = \mathbf{v}, \quad \mathbf{a}_{i+1} = \mathbf{a}_i - a_{i,i} \mathbf{e}_i + a_{i,i} \check{\mathbf{m}}_i = \mathbf{a}_i \mathbf{M}_i \quad i = 1, \dots, N. \quad (11)$$

Proof (1) Limit of $t_1^{(n)}/\theta^n$. By the definition of ν , which is a.s. finite, combined with the total workload process, we have, for $n > \nu$,

$$t_1^{(n)} = t^{(n)} = \sum + \sum_{i=1}^N \sum_{k=1}^{E_i(t^{(n)})} \bar{B}_i^{(k)} - \sum_{i=1}^N \sum_{k=1}^{X_i(t^{(n)})} \bar{B}_i^{(k)},$$

where $\bar{B}_i^{(k)}$ are i.i.d. copies of \bar{B}_i and $\sum = \sum_{l=0}^{\nu} (t_1^{(l)} - t^{(l)})$ is a.s. finite. Therefore, we have

$$t_1^{(n)} = t^{(n)} = \sum + t^{(n)} A_1^{(n)} - \theta^n A_2^{(n)},$$

where

$$A_1^{(n)} = \sum_{i=1}^N \frac{\sum_{k=1}^{E_i(t^{(n)})} \bar{B}_i^{(k)}}{E_i(t^{(n)})} \frac{E_i(t^{(n)})}{t^{(n)}},$$

$$A_2^{(n)} = \sum_{i=1}^N \frac{\sum_{k=1}^{X_i(t^{(n)})} \bar{B}_i^{(k)}}{X_i(t^{(n)})} \frac{X_i(t^{(n)})}{\theta^n},$$

then

$$\frac{t_1^{(n)}}{\theta^n} = \frac{t^{(n)}}{\theta^n} = \frac{A_2^{(n)} - \sum / \theta^n}{A_1^{(n)} - 1}. \quad (12)$$

By the SLLN and Proposition 4.1, we obtain, as $n \rightarrow \infty$,

$$A_1^{(n)} \rightarrow \sum_{i=1}^N \lambda_i \mathbb{E} \bar{B}_i^{(k)} = \sum_{i=1}^N \lambda_i \bar{c}_i = \rho, \quad A_2^{(n)} \rightarrow \sum_{i=1}^N v_i \xi \mathbb{E} \bar{B}_i^{(k)} = \sum_{i=1}^N v_i \bar{c}_i \xi \quad a.s.. \quad (13)$$

Then by (12) and (13), we have, as $n \rightarrow \infty$,

$$\frac{t^{(n)}}{\theta^n} \rightarrow b_1 \xi \quad \text{and} \quad \frac{t_1^{(n)}}{\theta^n} \rightarrow b_1 \xi,$$

where $b_1 = \frac{\sum_{i=1}^N v_i \bar{c}_i}{\rho - 1}$.

(2) **Limit of $t_i^{(n)}/\theta^n$.** In (1), by utilizing the index ν and equation $t_1^{(n)} = t^{(n)}$, we proved $\lim_{n \rightarrow \infty} t_1^{(n)}/\theta^n = b_1 \xi$. By the symmetry, there should also exist positive numbers b_i such that

$$\lim_{n \rightarrow \infty} \frac{t_i^{(n)}}{\theta^n} = b_i \xi, \quad i = 1, \dots, N.$$

It remains to prove these positive numbers b_i 's satisfy (9), which refer to (5) below.

(3) **Limit of $X(t_i)/\theta^n$ and (10).** By definition, we have

$$X_j(t_{i+1}^{(n)}) = \begin{cases} X_j(t_i^{(n)}) + E_j(t_{i+1}^{(n)}) - E_j(t_i^{(n)}) + \sum_{k=1}^{Y_i(I_i(t^{n+1})) - Y_i(I_i(t^n))} \delta_{\{A_i^{(k)}=j\}}, & j \neq i; \\ X_i(t_i^{(n)}) + E_i(t_{i+1}^{(n)}) - E_i(t_i^{(n)}) - \sum_{k=1}^{Y_i(I_i(t^{n+1})) - Y_i(I_i(t^n))} \delta_{\{A_i^{(k)} \neq i\}}, & j = i, \end{cases} \quad (14)$$

where $A_i^{(k)}$ are i.i.d. copies of A_i . Therefore, by SLLN and (14), we obtain

$$\frac{E_j(t_{i+1}^{(n)}) - E_j(t_i^{(n)})}{\theta^n} \rightarrow \lambda_j(b_{i+1} - b_i)\xi \quad a.s. \quad \text{as } n \rightarrow \infty,$$

and

$$\begin{aligned} & \frac{\sum_{k=1}^{Y_i(I_i(t^{n+1})) - Y_i(I_i(t^n))} \delta_{\{A_i^{(k)}=j\}}}{\theta^n} \\ &= \frac{\sum_{k=1}^{Y_i(I_i(t^{n+1})) - Y_i(I_i(t^n))} \delta_{\{A_i^{(k)}=j\}}}{Y_i(I_i(t^{n+1})) - Y_i(I_i(t^n))} \frac{Y_i(I_i(t^{n+1})) - Y_i(I_i(t^n))}{I_i(t^{n+1}) - I_i(t^n)} \frac{I_i(t^{n+1}) - I_i(t^n)}{\theta^n} \\ &\rightarrow \mu_i(b_{i+1} - b_i)\xi \mathbb{E}\delta_{\{A_i^{(k)}=j\}} = p_{i,j}\mu_i(b_{i+1} - b_i)\xi. \end{aligned}$$

Similarly,

$$\frac{\sum_{k=1}^{Y_i(I_i(t^{n+1})) - Y_i(I_i(t^n))} \delta_{\{A_i^{(k)} \neq i\}}}{\theta^n} \rightarrow (1 - p_{i,i})\mu_i(b_{i+1} - b_i)\xi.$$

It follows that

$$\frac{X_j(t_{i+1}^{(n)})}{\theta^n} \rightarrow a_{i+1,j}\xi,$$

where

$$a_{i+1,j} = \begin{cases} a_{i,j} + \lambda_j(b_{i+1} - b_i) + p_{i,j}\mu_j(b_{i+1} - b_i), & j \neq i; \\ a_{i,i} + \lambda_i(b_{i+1} - b_i) - \mu_i(1 - p_{i,i})(b_{i+1} - b_i), & j = i. \end{cases}$$

(4) **Equation (11).** (11) follows from the below equation

$$\mathbf{Q}(t_{i+1}^{(n)}) = \mathbf{Q}(t_i^{(n)}) - Q_i(t_i^{(n)})\mathbf{e}_i + \sum_{k=1}^{Q_i(t_i^{(n)})} \check{\mathbf{L}}_i^{(n,k)},$$

where $\check{\mathbf{L}}_i^{(n,k)} = (\check{L}_{i,1}^{(n,k)}, \dots, \check{L}_{i,N}^{(n,k)})$ and $\check{L}_{i,j}^{(n,k)}$ are i.i.d. copies of $\check{L}_{i,j}$.

(5) **Equation (9).** Recall that

$$t_{i+1}^{(n)} = t_i^{(n)} + \sum_{k=1}^{X_i(t_i^{(n)})} T_i^{(k)}, \quad (15)$$

$$X_i(t_i^{(n)}) = X_i(t_1^{(n)}) + E_i(t_i^{(n)}) - E_i(t_1^{(n)}) + \sum_{j=1}^{i-1} \sum_{k=1}^{Y_j(I_j(t^{(n)})) - Y_j(I_j(t^{(n-1)}))} \delta_{\{A_j^{(k)}=i\}}. \quad (16)$$

From Lemma 7.1 and (15), we get

$$b_{i+1} - b_i = a_{i,i} t_i. \quad (17)$$

By (16) and the SLLN, we obtain

$$a_{i,i} = v_i + \lambda_i(b_i - b_1) + \sum_{j=1}^{i-1} \mu_j(b_{j+1} - b_j) p_{j,i}. \quad (18)$$

Then (9) can be proved by substituting (18) into (17).

By $t_{N+1}^{(n)} = t_1^{(n+1)}$, we obtain $b_{N+1}\xi = \lim_{n \rightarrow \infty} t_{N+1}^{(n)}/\theta^n = \lim_{n \rightarrow \infty} \theta t_1^{(n+1)}/\theta^{n+1} = \theta b_1 \xi$, i.e. $b_{N+1} = \theta b_1$. This can be proved as follows. By the definition of \mathbf{M}_i in Lemma 4.2, it is easy to give

$$(\mathbf{M}_i - \mathbf{I})\bar{\mathbf{c}}^T = (\rho - 1)t_i \mathbf{e}_i,$$

where \mathbf{I} is the identity matrix and $\bar{\mathbf{c}} = (\bar{c}_1, \dots, \bar{c}_N)$. Substituting the above equation into (17) yields

$$b_{i+1} - b_i = a_{i,i} t_i = \mathbf{a}_i \mathbf{e}_i t_i = \frac{1}{\rho - 1} \mathbf{a}_i (\mathbf{M}_i - \mathbf{I}) \bar{\mathbf{c}}^T = \frac{1}{\rho - 1} (\mathbf{a}_{i+1} - \mathbf{a}_i) \bar{\mathbf{c}}^T.$$

which gives immediately

$$\begin{aligned} b_{N+1} &= \sum_{i=1}^N (b_{i+1} - b_i) + b_1 = \frac{1}{\rho - 1} (\mathbf{a}_{N+1} - \mathbf{a}_1) \bar{\mathbf{c}}^T + b_1 \\ &= \frac{1}{\rho - 1} (\mathbf{v} \mathbf{M} - \mathbf{v}) \bar{\mathbf{c}}^T + b_1 = (\theta - 1) \frac{\mathbf{v} \bar{\mathbf{c}}^T}{\rho - 1} + b_1 = (\theta - 1) b_1 + b_1 = \theta b_1. \end{aligned}$$

Proof of Theorem 5.1. With the results of Lemma 7.2, Theorem 5.1 immediately follows from Lemma 6 in [15].

Proof of Theorem 5.2. For each i , by (5), we know that the function $\bar{X}_i(\cdot)$ might have discontinuities only at $t = 0$ and $t = \theta^k \bar{b}_i$ for each $k \in \mathbb{N}$. Since the function $\bar{X}_i(\cdot)$ is càdlàg, the continuity of $\bar{X}(\cdot)$ is evident in combination with the definition of \mathbf{a}_i .

Additionally, the uniform convergence on compact sets can be proved in the same way as in the proof of Theorem 2 in [15]. Hence, it suffices to prove the point-wise convergence (5) for each $i = 1, 2, \dots, N$.

By Theorem 5.1, we have, as $n \rightarrow \infty$,

$$\begin{aligned} \frac{t_i^{(\eta_n+k)}}{\theta^n} &\rightarrow \theta^k \bar{b}_i \xi \quad \text{and} \quad \frac{\mathbf{X}(t_i^{(\eta_n+k)})}{\theta^n} \rightarrow \xi \theta^k \bar{\mathbf{a}}_i, \\ \frac{I_i(t_i^{(\eta_n+k)})}{\theta^n} &\rightarrow \theta^k \frac{\bar{b}_{i+1} - \bar{b}_i}{\theta(\theta - 1)} \xi, \\ \frac{E_i(t)}{t} &\rightarrow \lambda_i \quad \text{and} \quad \frac{Y_i(t)}{t} \rightarrow \mu_i. \end{aligned}$$

We will show that, as $n \rightarrow \infty$,

$$\bar{\mathbf{X}}^{(n)}(t) \rightarrow \xi \bar{\mathbf{X}}\left(\frac{t}{\xi}\right) \quad \text{for all } t \in [0, \infty), \quad (19)$$

where $\bar{\mathbf{X}}^{(n)}(\cdot)$ is given by the form of (4).

For $t = 0$, the convergence of (19) holds since the system starts empty. For each $i = 1, \dots, N$, if $t \in [\theta^k \bar{b}_i, \theta^k \bar{b}_{i+1})$, it remains to prove

$$\bar{X}_j(t) = \begin{cases} \theta^k \bar{a}_{i,i} + [\lambda_i - \mu_i(1 - p_{i,i})](t - \theta^k \bar{b}_i), & j = i; \\ \theta^k \bar{a}_{i,j} + [\lambda_j + \mu_i p_{i,j}](t - \theta^k \bar{b}_i), & j \neq i. \end{cases}$$

For all n big enough, $\frac{t_i^{(\eta_n+k)}}{\theta^n} < t < \frac{t_{i+1}^{(\eta_n+k)}}{\theta^n}$ implying that Q_i is in service during $[t_i^{(\eta_n+k)}, \theta^n t)$. Hence,

$$X_j(\theta^n t) = \begin{cases} X_i(t_i^{(\eta_n+k)}) + E_i(\theta^n t) - E_i(t_i^{(\eta_n+k)}) - \sum_{k=1}^{Y_i(I_i(\theta^n t)) - Y_i(I_i(t_i^{(\eta_n+k)}))} \delta_{\{A_i^{(k)} \neq i\}}, & j = i; \\ X_j(t_i^{(\eta_n+k)}) + E_j(\theta^n t) - E_j(t_i^{(\eta_n+k)}) + \sum_{l=1}^{Y_i(I_i(\theta^n t)) - Y_i(I_i(t_i^{(\eta_n+k)}))} \delta_{\{A_i^{(l)} = j\}}, & j \neq i. \end{cases}$$

Therefore,

$$\bar{X}_j^{(n)}(t) = \frac{X_j(\theta^n t)}{\theta^n} \rightarrow \begin{cases} \xi \theta^k \bar{a}_{i,i} + [\lambda_i - \mu_i(1 - p_{i,i})](t - \theta^k \bar{b}_i \xi), & j = i; \\ \xi \theta^k \bar{a}_{i,j} + (\lambda_j + \mu_i p_{i,j})(t - \theta^k \bar{b}_i \xi), & j \neq i, \end{cases}$$

where the right hand-side actually equals $\xi \bar{X}_j(\frac{t}{\xi})$. And the proof is concluded.

8. Numerical validation and optimization of gating indexes

8.1. Numerical validation

Tab. 1 Parameter values in 3-queue polling network

Parameter	Considered parameter values
Arriving rate	$\lambda_1 = \lambda_2 = \lambda_3 = 1$
Service rate	$\mu_1 = 8, \mu_2 = 5, \mu_3 = 2$
Transition probability	$P = (p_{i,j})_{3 \times 3} = \begin{pmatrix} 0.1 & 0.25 & 0.2 \\ 0.2 & 0.1 & 0.2 \\ 0.2 & 0.1 & 0.25 \end{pmatrix}$

This subsection is devoted to test the validity of the fluid limits of the scaled queue length processes defined in (4). For simplicity, we consider a 3-queue polling system described in Tab.1 with exponentially distributed service times. For this model, it is readily to obtain $\rho_1 = 0.4749$, $\rho_2 = 0.5194$, $\rho_3 = 0.8625$ and $\rho = 1.8568$, which belongs to the overloaded traffic case studied in this paper.

We utilize the SimEvents toolbox of Matlab to undertake the simulations of the polling networks. The exhaustive and gated service policies are taken for example and some vital variables are given in Tab.2. In order to illustrate the convergence of the scaled queue length processes defined in (4), we take $n = 1, 5, 8, 10$ in polling network with exhaustive service policies and $n = 1, 10, 18, 20$ in the gated counterpart. The corresponding scaled queue length processes at Q_2 and the scaled total queue length process are depicted in Fig.3 and Fig.4, respectively. Apparently, the scaled queue length sample paths get closer and closer as n increases.

Moreover, as shown in Fig.3 and Fig.4, the fluid limit processes both demonstrate an oscillation waveform with increasing amplitude and cycle time forward and oscillate at an infinite rate when approaching zero. According to Theorem 5.2, the amplitude and cycle time increase by $\theta - 1$ times each cycle, which has been easily verified by the sample paths.

8.2. Optimization of gating indexes

Subsequently, we consider the optimization of the gating indexes by numerical method. It is assumed that the gating indexes are integers additionally. Virtually, the fluid limits only depend on the exhaustiveness of the

Tab. 2 Essential variable values in 3-queue polling network

Variable	values	
	Exhaustive	Gate
Gating index	∞	1
Exhaustiveness	1	$1 - (\frac{\lambda_i}{\mu_i} + p_{i,i})$
Maximum eigenvalue	$\theta = 3.7497$	$\theta = 1.6394$
Left eigenvector	$v = [0.9731, 0.683, 0]$	$v = [0.7454, 0.5301, 0.4774]$

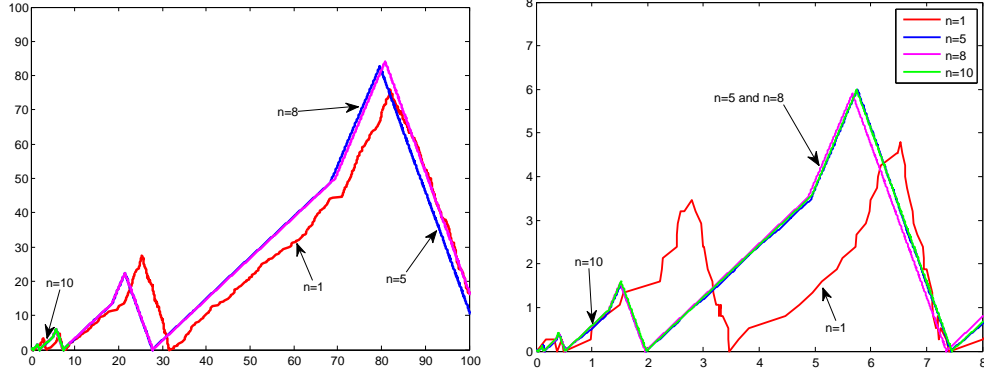


Fig. 3 The scaled queue length process at Q_2 for different n (left: $t \in [0, 100]$, right: $t \in [0, 8]$)

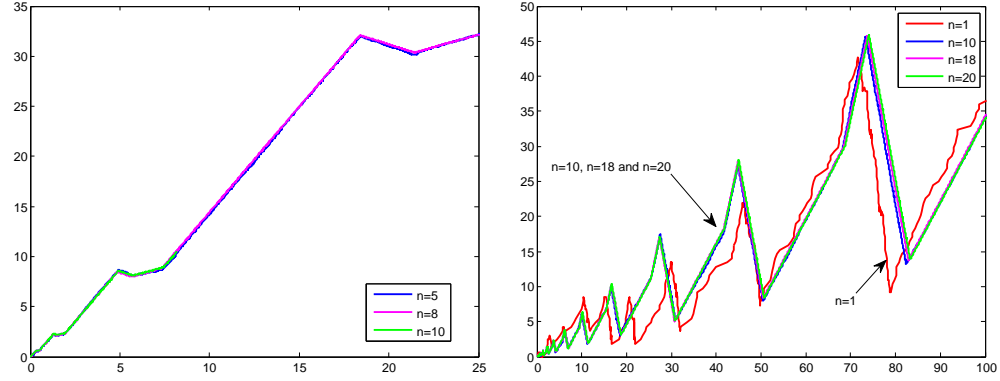


Fig. 4 Left: the scaled total queue length process for different n with exhaustive service policy
Right: the scaled queue length process at Q_2 for different n with gated service policy

service discipline at each queue (moment of gating index), which allows us to minimize the total queue length process through the accommodation of the integer gating indexes.

By (6), the average growth rate of the total queue length process β with exhaustive and gated service policies equals to 1.5025 and 1.2416 respectively (see Fig.5). This can be intuitively interpreted from the growth rate during each visiting period on different queues. The visiting period at Q_3 with the maximal growth rate (minimal service rate) takes 4 times as much time as others in exhaustive service policy. Instead, it takes less than 2 times as much time as others in gated service policy. Therefore, to minimize the average growth rate, we need to increase the visiting time at Q_1 and Q_2 and decrease the visiting time at Q_3 .

To optimize the gating indexes turns to be an integer programming with three variables. The GA toolbox of Matlab is undertaken here to search for the optimal gating indexes. For our model, the GA solver just takes 51 iterations to find the optimal solution: Q_1 and Q_2 both take exhaustive service policy while Q_3 takes gated service policy. The minimal average growth rate equals to 1.19262 and the corresponding exhaustiveness is $f_1 = f_2 = 1$, $f_3 = 0.25$. Fig.5 depicts the process of the optimal average growth rate in each generation. Apparently, the convergence process turns to be very effective. Hence, the average growth rate of the fluid limit provides a simple and transparent method to optimize the gating index.

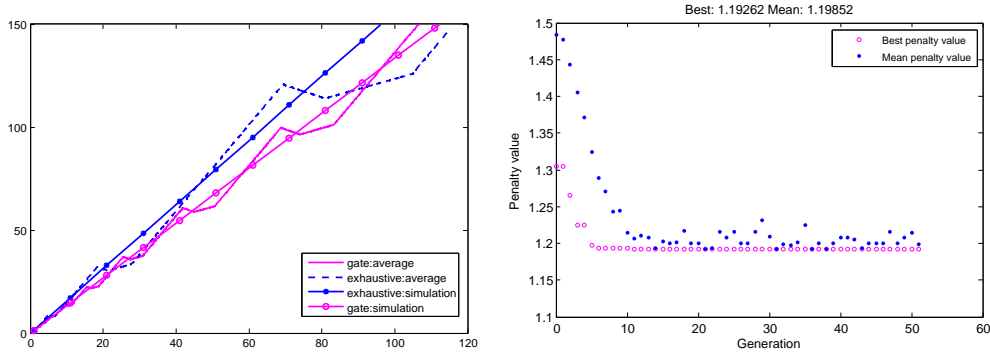


Fig. 5 Left: the fluid limit of the scaled total queue length process (exhaustive and gate)
Right: the convergence process of the optimal average growth rate by GA solver

9. Conclusions and Further Research

Inspired by [15], we present the fluid limit of an overloaded polling system with general random multi-gated service discipline and customer re-routing policies. These results provide new fundamental insight in the impact of exhaustiveness. As a by-product, we propose an optimization problem of gating indexes to minimize the total queue length process.

This work gives rise to a variety of directions for further research. A logical follow-up step would be to study the case with non-zero switch-over time and more general branching-type polling models. In addition, the asymptotic behaviors of discrete-time polling systems are also direct extensions to this study. Furthermore, the fluid limit allows us to propose control strategies of the growth depression, which requires substantially more effort.

Acknowledgement

This research is partially supported by the National Natural Science Foundation of China (11671404, 11571052), by the Fundamental Research Funds for the Central Universities of Central South University(2015zzts012), and by the Fundamental Research Funds for the Central Universities(WUT:2017 IVA 069).

References

- [1] M. A. Boon, R. Van der Mei, E. Winands, Applications of polling systems, *Surv. Oper. Res. Manag. Sci.* 16 (2) (2011) 67–82.
- [2] H. Levy, M. Sidi, Polling systems: applications, modeling, and optimization, *IEEE Trans. Commun.* 38 (10) (1990) 1750–1760.
- [3] M. A. Boon, R. D. van der Mei, E. M. Winands, Queueing networks with a single shared server: light and heavy traffic, *ACM SIGMETRICS Perform. Eval. Rev.* 39 (2) (2011) 44–46.
- [4] M. A. Boon, R. D. van der Mei, E. M. Winands, Waiting times in queueing networks with a single shared server, *Queueing Syst.* 74 (4) (2013) 403–429.
- [5] M. Sidi, H. Levy, Customer routing on polling systems, in: *Performance 1990*, eds. P.J.B. King, I. Mitrani and R.J. Pooley (North-Holland, Amsterdam, 1990) pp. 319–331.

- [6] M. Sidi, H. Levy, S. W. Fuhrmann, A queueing network with a single cyclically roving server, *Queueing Syst.* 11 (1) (1992) 121–144.
- [7] E. Coffman Jr, A. Puhalskii, M. Reiman, Polling systems with zero switchover times: a heavy-traffic averaging principle, *Ann. Appl. Probab.* (1995) 681–719.
- [8] E. Coffman Jr, A. Puhalskii, M. Reiman, Polling systems in heavy traffic: A Bessel process limit, *Math. Oper. Res.* 23 (2) (1998) 257–304.
- [9] R. D. van der Mei, Towards a unifying theory on branching-type polling systems in heavy traffic, *Queueing Syst.* 57 (1) (2007) 29–46.
- [10] Z. Liu, Y. Chu, J. Wu, The asymptotic behavior of a branching-type polling network in heavy traffic, *Sci.China-Math. (Chinese Series)* 45 (5) (2015) 515–526.
- [11] Z. Liu, Y. Chu, J. Wu, Heavy-traffic Asymptotics of Priority Polling System with Threshold Service Policy, *Comput. Oper. Res.* 65 (2016) 19–28.
- [12] A. L. Puha, A. L. Stolyar, R. J. Williams, The fluid limit of an overloaded processor sharing queue, *Math. Oper. Res.* 31 (31) (2006) 316–350.
- [13] R. Talreja, W. Whitt, Fluid models for overloaded multiclass many-server queueing systems with first-come, first-served routing, *Manage. Sci.* 54 (8) (2008) 1513–1527.
- [14] O. B. Jennings, J. E. Reed, An overloaded multiclass FIFO queue with abandonments, *Oper. Res.* 60 (5) (2012) pgs. 1282–1295.
- [15] M. Remerova, S. F. B. Zwart, Random fluid limit of an overloaded polling model, *Adv. in Appl. Probab.* 46 (1) (2013) 76–101.
- [16] K. B. Athreya, P. E. Ney, *Branching Processes*, Springer Berlin Heidelberg, 1972.
- [17] J. Resing, Polling systems and multitype branching processes, *Queueing Syst.* 13 (4) (1993) 409–426.
- [18] R. D. van der Mei, Polling systems with switch-over times under heavy load: moments of the delay, *Queueing Syst.* 36 (4) (2000) 381–404.